Chapitre 1

La régression linéaire simple

Si l'état décide d'accentuer les mesures contre l'alcool au volant, quel en sera l'effet sur la sécurité routière ? Réduire la taille des classes à l'école primaire aura-t-il des conséquences quantifiables sur les performances scolaires ? Quel sera l'impact du prolongement d'une ou de plusieurs années de votre parcours universitaire sur votre revenu futur ?

Ces trois questions portent sur l'évaluation de l'effet de la variation d'une variable X (les mesures contre l'alcool au volant, l'effectif de la classe, le prolongement du parcours universitaire) sur une variable Y (la sécurité routière, la performance scolaire, le revenu).

Ce chapitre introduit les modèles de régression linéaire reliant une variable X à une autre variable Y. Dans ces modèles, la pente de la droite reliant X et Y (la droite de régression) représente l'effet sur Y de la variation d'une unité (ou variation unitaire) de X. Tout comme la moyenne de Y, la pente de la droite de régression est une caractéristique inconnue de la distribution jointe de X et Y. L'objet de l'économétrie est d'estimer cette pente, c'est-à-dire d'estimer l'effet d'une variation unitaire de X sur Y à partir des observations des deux variables.

Ce chapitre décrit les méthodes d'estimation de cette pente. Par exemple, en utilisant les données sur la performance scolaire associée à l'effectif des classes dans différentes écoles, nous montrons comment estimer l'effet d'une réduction d'effectif sur la performance scolaire. Nous estimerons la pente et la constante de la droite de régression reliant X à Y par la méthode des moindres carrés ordinaires (MCO).

1.1 Le modèle de régression linéaire

Le directeur administratif d'une zone scolaire vous demande conseil pour établir sa politique de recrutement des professeurs à l'école primaire. Il aimerait réduire le nombre d'élèves par enseignant (le ratio élèves-enseignants) de deux élèves. Mais il se trouve devant un dilemme : d'une part, les parents souhaitent que leurs enfants soient bien suivis, bien encadrés et dans des structures éducatives appropriées ; d'autre part, les collectivités locales se refusent, pour des considérations budgétaires, à financer le recrutement de nouveaux professeurs. Pour arbitrer, le directeur doit évaluer l'effet de la réduction des classes sur la performance scolaire.

Dans certaines écoles américaines, la performance scolaire est mesurée par des tests standard. La promotion, ainsi que la rémunération, de certains administrateurs sont en partie tributaires des scores de ces tests. De ce fait, la question du directeur peut être reformulée comme suit : quel sera l'effet d'une réduction de deux élèves par classe sur la performance scolaire ?

Répondre précisément à cette question nécessite une évaluation quantitative de l'effet de cette réduction. On peut la formuler mathématiquement comme suit :

$$\beta_1 = \frac{\textit{variation de ScoreTest}}{\textit{variation de TailleClasse}} = \frac{\Delta \textit{ScoreTest}}{\Delta \textit{TailleClasse}}, \tag{1.1}$$

où β_1 désigne la variation du score des tests lorsqu'on modifie la taille des classes d'une unité. L'équation (1.1) peut s'écrire sous la forme

$$\Delta \textit{ScoreTest} = \beta_1 \times \Delta \textit{TailleClasse}. \tag{1.2}$$

Estimer β_1 nous permet de quantifier l'effet d'une variation des effectifs sur le score des tests. Par exemple, si $\beta_1 = -0{,}6$, une réduction des classes de deux élèves correspond à une variation estimée de $(-0{,}6) \times (-2) = 1{,}2$ du score des tests. Autrement dit, l'économètre prédit une hausse du score des tests de 1,2 point si l'on réduit les classes de deux élèves.

L'équation (1.1) correspond à la définition de la pente d'une droite de régression reliant le score des tests à l'effectif des classes. Cette droite peut être écrite sous la forme

$$\textit{ScoreTest} = \beta_0 + \beta_1 \times \textit{TailleClasse}, \tag{1.3}$$

où β_0 et β_1 sont respectivement la constante et la pente de la droite. Selon l'équation (1.3), la connaissance des valeurs de β_0 et β_1 permet non seulement de déterminer, pour une zone scolaire donnée, la variation du score des tests associée à un changement de l'effectif des classes, mais aussi de prédire la valeur du score moyen pour une valeur donnée de la taille des classes.

Cependant, la réponse qui sera proposée au directeur sur la base de l'équation (1.3) ne va sûrement pas le satisfaire. En effet, il pensera, à juste titre, que l'effectif des classes n'est qu'un des facteurs susceptibles d'affecter la performance scolaire. Ainsi, deux zones scolaires disposant de classes de même taille peuvent avoir, pour de multiples raisons, des performances scolaires différentes. Par exemple, une zone scolaire peut se différencier par la qualité de ses

professeurs et de ses programmes pédagogiques. Même si deux zones scolaires ont des effectifs par classe et un environnement pédagogique comparables, elles peuvent différer par leurs populations d'élèves : une zone peut contenir par exemple plus de primo-arrivants (et donc moins d'élèves maîtrisant la langue d'enseignement) ou plus de familles aisées. Finalement, ce même directeur mettra en exergue le rôle des facteurs aléatoires susceptibles d'altérer, le jour du test, les performances des élèves et d'ainsi générer des scores de tests différents pour deux zones possédant les mêmes caractéristiques. Pour toutes ces raisons, l'équation (1.3) ne décrit pas de manière exacte l'effet dans chaque zone. Elle peut être considérée comme la formulation d'un effet moyen, si l'on considère l'*ensemble* (ou la « population ») des zones scolaires.

Une version de cette relation linéaire fonctionnant pour *chaque* zone scolaire doit incorporer d'autres facteurs influençant les scores des tests, incluant les caractéristiques propres à chaque zone (la qualité de leurs enseignants, la composition de leurs élèves, les conditions du déroulement des tests, etc.). Dans le chapitre 3, on s'intéressera plus particulièrement à une approche qui permet d'établir une liste des facteurs les plus prépondérants et de les introduire explicitement dans l'équation (1.3). Pour simplifier, nous les regroupons dans « autres facteurs » et nous réécrivons, pour une zone scolaire donnée, l'équation (1.3) :

$$ScoreTest = \beta_0 + \beta_1 \times TailleClasse + \text{autres facteurs}. \tag{1.4}$$

Ainsi, le score des tests pour une zone donnée est exprimé en termes d'une première composante, $\beta_1 \times TailleClasse$, représentant l'effet moyen de l'effectif des classes sur les scores de l'ensemble des zones scolaires, et d'une deuxième composante comportant les autres facteurs.

Bien que cette discussion soit focalisée sur l'étude d'une relation spécifique liant le score des tests à l'effectif des classes, il est intéressant de replacer cette étude dans un contexte plus général. Pour cela, supposons que nous disposons de n zones et que Y_i et X_i sont les moyennes respectives des scores des tests et de l'effectif des classes pour la $i^{ème}$ zone. Le terme aléatoire u_i désigne les autres facteurs influençant le score des tests de la $i^{ème}$ zone. Ainsi, l'équation (1.4) peut s'écrire sous la forme

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \tag{1.5}$$

où pour chaque zone i ($i = 1, \ldots, n$), β_0 et β_1 sont respectivement la constante et la pente de cette droite de régression. L'équation (1.5) est la représentation d'un **modèle de régression linéaire avec un seul régresseur (modèle de régression simple)** dont Y et X sont respectivement la **variable dépendante** et la **variable indépendante (régresseur)**. Le premier membre de l'équation (1.5), $\beta_0 + \beta_1 X_i$, est la **droite de régression théorique** ou la **fonction de régression théorique**. Celle-ci représente la relation entre Y et X moyennée sur l'ensemble de la population. Ainsi, sur la base de cette droite de régression, la connaissance de la valeur de X permettra de prédire la valeur de la variable dépendante, Y, donnée par $\beta_0 + \beta_1 X$.

La **constante** β_0 et la **pente** β_1 sont des coefficients ou des **paramètres** de la droite de régression. La pente correspond à la variation de Y résultant d'une variation unitaire de X ; la constante est la valeur de la droite de régression pour $X = 0$. C'est le point d'intersection entre la droite de régression et l'axe des Y. Dans certaines applications économétriques, la constante a une interprétation économique pertinente, alors que dans d'autres, elle n'a aucune signification. À titre d'exemple, si X correspond à l'effectif des classes, interpréter la constante comme la valeur estimée du score des tests pour une classe vide n'a aucun sens. Elle peut être, cependant, interprétée d'une manière mathématique comme étant le paramètre qui détermine le niveau de la droite de régression.

Dans l'équation (1.5), le terme u_i est le **terme d'erreur** ; il incorpore tous les autres facteurs expliquant la différence entre la moyenne du score des tests de la $i^{ème}$ zone et la valeur estimée par la droite de régression. Ce terme d'erreur comporte, en dehors de X, tous les autres facteurs qui déterminent, pour une observation i donnée, la valeur de la variable dépendante Y. Dans l'exemple portant sur l'effectif des classes, ces autres facteurs comportent toutes les caractéristiques propres à la $i^{ème}$ zone scolaire (les qualités des professeurs, la composition des élèves, les circonstances de l'examen, les éventuelles erreurs d'évaluation, etc.) et qui affectent la performance des élèves durant la période du déroulement des tests.

Le modèle de régression linéaire et sa terminologie sont résumés dans le concept-clé 1.1.

■ Concept-clé 1.1 Terminologie pour le modèle de régression linéaire simple

Le modèle de régression linéaire est

$$Y_i = \beta_0 + \beta_1 X_i + u_i,$$

où

- l'indice i désigne les observations individuelles, $i = 1, \ldots, n$;
- Y_i est la variable dépendante ;
- X_i est la variable indépendante, le régresseur ;
- $\beta_0 + \beta_1 X_i$ est la droite de régression théorique ou la fonction de régression théorique ;
- β_0 est la constante de la droite de régression théorique ;
- β_1 est la pente de la droite de régression théorique ;
- u_i est le terme d'erreur.

La figure 1.1 montre la droite de régression linéaire simple pour sept observations hypothétiques du score des tests (Y) et de l'effectif des classes (X). Cette droite de régression théorique est donnée par $\beta_0 + \beta_1 X$. Une pente négative de

la droite de régression ($\beta_1 < 0$) signifie qu'une zone scolaire avec un faible ratio élèves-enseignants (faibles effectifs des classes) a tendance à avoir des scores de tests élevés. La constante s'interprète comme le point d'intersection entre l'axe des Y et la droite de régression linéaire. Comme nous l'avons déjà mentionné plus haut, elle n'a pas de signification pertinente.

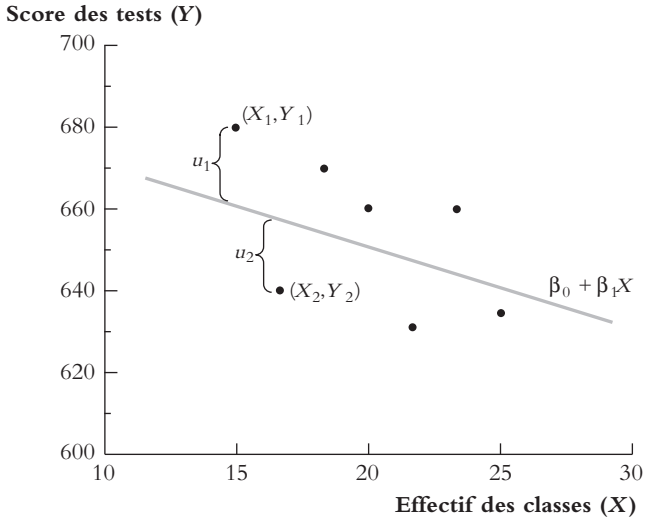

Figure 1.1 Score des tests et ratio élèves-enseignants (données hypothétiques)

Le nuage de points représente des observations hypothétiques pour sept zones scolaires. La droite de régression théorique est donnée par $\beta_0 + \beta_1 X$. La distance séparant le $i^{ème}$ point de la droite de régression théorique est $Y_i - (\beta_0 + \beta_1 X_i)$; elle correspond au terme d'erreur u_i associé à la $i^{ème}$ observation.

La différence entre les observations hypothétiques et la droite de régression linéaire, illustrée par la figure 1.1, est expliquée par la présence d'autres facteurs qui déterminent la performance des tests et qui sont contenus dans le terme d'erreur. À titre d'exemple, pour la zone scolaire 1, un écart positif entre la valeur Y_1 et sa valeur estimée (obtenue à partir de la droite de régression) signifie que le score des tests dans la zone 1 a été sous-estimé, c'est-à-dire que le terme d'erreur pour cette zone est positif. Pour la zone 2, nous obtenons la configuration contraire : la valeur effectivement observée de Y_2 est au-dessous de cette même droite de régression, ce qui signifie que, pour cette zone, le score des tests a été surestimé et que par conséquent $u_2 < 0$.

Revenons maintenant à la problématique soulevée par le directeur et formulée par la question suivante : quel sera l'effet attendu d'une réduction de l'effectif des classes de deux élèves par enseignant sur le score des tests ? La réponse est simple : la variation attendue est de $(-2) \times \beta_1$. Mais quelle est la valeur de β_1 ?

1.2 Estimation des paramètres du modèle de régression linéaire

Dans la pratique, les coefficients de la constante et de la pente, respectivement notés β_0 et β_1, qui lient le score des tests à l'effectif des classes, sont inconnus. Leur estimation passe par l'utilisation d'un échantillon d'observations correspondant à ces deux variables.

Le problème d'estimation de ces deux paramètres est similaire au problème soulevé lors de l'estimation de la moyenne empirique ($\bar{Y}$) de l'annexe B. À titre d'exemple, supposons qu'on souhaite comparer les salaires moyens des hommes et des femmes récemment diplômés de l'université. Bien que les moyennes théoriques des rémunérations des deux sexes soient inconnues, il est possible de les estimer à partir d'un échantillon aléatoire composé des jeunes diplômés de l'université. Ainsi, pour les femmes, l'estimateur naturel de cette moyenne théorique (inconnue) est égal à la moyenne des salaires des femmes diplômées de l'université dans l'échantillon.

La même idée s'étend au modèle de régression linéaire. En effet, bien que la valeur de la pente, β_1, de la droite de régression reliant le score des tests à l'effectif de la classe soit inconnue, il est possible de l'estimer en utilisant l'échantillon des données.

Prenons un échantillon d'étude comportant les observations sur les scores des tests et les effectifs par classe en 1999, pour 420 zones scolaires de Californie. Le score des tests est la moyenne des scores en lecture et en mathématiques pour les cinq premières années d'études. L'effectif d'une classe peut être mesuré de plusieurs manières. Ici, nous allons utiliser le nombre d'élèves dans la zone scolaire divisé par le nombre d'enseignants. Cette mesure, exprimée en termes de ratio élèves-enseignants, est la plus utilisée dans la pratique ; elle sera adoptée dans la suite du chapitre.

Le tableau 1.1 résume les distributions des scores des tests et des effectifs par classe de notre échantillon d'observations. La moyenne et l'écart-type du ratio élèves-enseignants sont respectivement égaux à 19,6 et à 1,9 élèves par enseignant. Le $10^{\text{ème}}$ percentile de la distribution de ce ratio est de 17,3 (c'est-à-dire que 10 % des zones ont un ratio élèves-professeur au-dessous de 17,3), alors que la zone appartenant au $90^{\text{ème}}$ percentile admet un ratio de 21,9.

Tableau 1.1 Synthèse de la distribution des ratios élèves-enseignants et des scores des tests

			Percentile						
	Moy.	Écart-type	10 %	25 %	40 %	50 %	60 %	75 %	90 %
REE	19,6	1,9	17,3	18,6	19,3	19,7	20,1	20,9	21,9
Score	654,2	19,1	630,4	640,0	649,1	654,5	659,4	666,7	679,1

La figure 1.2 illustre la relation liant le score des tests à l'effectif des classes. On trouve une valeur de corrélation empirique de −0,23, ce qui indique l'existence d'une faible relation négative entre ces deux variables. Bien qu'on associe les faibles scores aux classes de grands effectifs, il existe, cependant, d'autres déterminants de la variable dépendante (les scores des tests) générant des observations du nuage des points en dessous ou au-dessus de la droite de régression. En dépit de cette faible corrélation, il est possible de construire une droite dont la valeur inconnue de la pente est à estimer. La méthode la plus utilisée est celle des moindres carrés ordinaires (MCO).

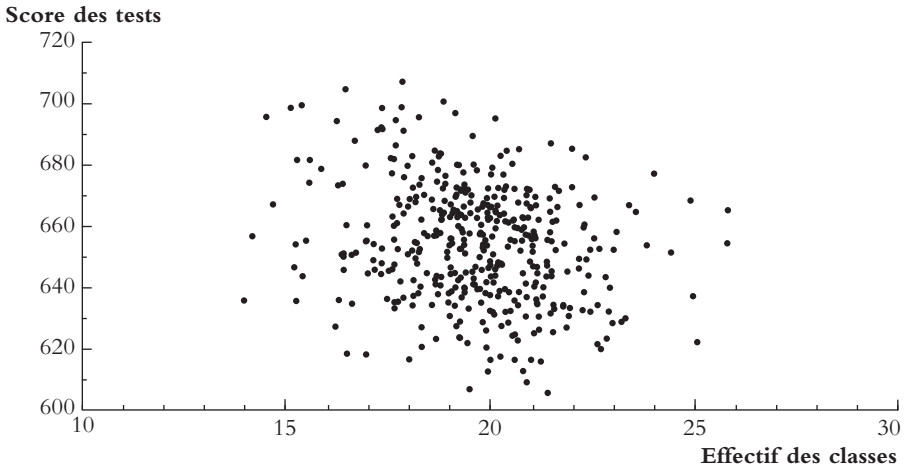

Figure 1.2 Score des tests et ratio élèves-enseignants (données réelles)

Données pour 420 zones scolaires de Californie. Il existe une légère relation négative entre le score des tests et le ratio élèves-enseignants : la corrélation empirique est égale à −0,23.

1.2.1 Méthode des moindres carrés ordinaires

La méthode des MCO permet d'estimer les coefficients de régression de manière à ce que la droite de régression estimée soit la plus proche possible des données observées. Cela revient à supposer que, dans la prédiction de Y sachant X (la définition de Y sachant X est rappelée en annexe A), la somme des carrés des erreurs est la plus faible possible.

Nous rappelons dans l'annexe B que la moyenne empirique, $\bar{Y}$, est l'estimateur des MCO de la moyenne théorique $E(Y)$; c'est-à-dire que parmi tous les estimateurs possibles, la valeur de $\hat{m} = \bar{Y}$ correspond à la plus petite valeur de la somme des carrés des erreurs, $\sum_{i=1}^{n}(Y_i - m)^2$.

Cette propriété caractérisant les moindres carrés ordinaires s'applique aussi aux paramètres d'un modèle de régression linéaire. Supposons que b_0 et b_1 sont deux estimateurs potentiels de β_0 et β_1. La droite de régression fondée sur ces

estimateurs est $b_0 + b_1X$ et la valeur estimée de Y_i est par conséquent donnée par $b_0 + b_1X_i$. Ainsi l'erreur de prédiction associée à la $i^{ème}$ observation est $Y_i - b_0 - b_1X_i$. Pour les n observations de notre échantillon, la somme des carrés de ces erreurs de prédiction est donnée par

$$\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2. \tag{1.6}$$

Pour le modèle de régression linéaire, la somme des carrés des résidus, donnée par l'équation (1.6), est une extension de l'expression de la somme des carrés des écarts correspondant à l'estimation de la moyenne (voir section 1, annexe B). De même qu'il existe un unique estimateur $\bar{Y}$ qui minimise $\sum_{i=1}^{n}(Y_i - m)^2$ de β_0 et β_1 qui minimisent l'expression (1.6). Ces estimateurs sont appelés les **estimateurs des moindres carrés ordinaires (MCO)** de β_0 et β_1.

Les estimateurs des MCO de β_0 et β_1 sont respectivement notés $\hat{\beta}_0$ et $\hat{\beta}_1$. La droite de régression des MCO, appelée aussi la **droite de régression empirique** ou la **fonction de régression empirique**, est une droite construite à partir des estimateurs des MCO : $\hat{\beta}_0 + \hat{\beta}_1X$. Cette dernière expression correspond aussi à la **valeur estimée** de Y_i. Le terme résiduel pour la $i^{ème}$ observation est la différence entre Y_i et sa valeur estimée : $\hat{u}_i = Y_i - \hat{Y}_i$. Les estimateurs par les MCO, $\hat{\beta}_0$ et $\hat{\beta}_1$, constituent l'équivalent empirique des paramètres théoriques β_0 et β_1. De même, la droite de régression des MCO, $\hat{\beta}_0 + \hat{\beta}_1X$, est la contrepartie empirique de la droite de régression théorique $\beta_0 + \beta_1X$. Enfin, l'estimateur des MCO, $\hat{u}_i$, est la contrepartie empirique du terme d'erreur théorique u_i. Il est possible de calculer les estimateurs des MCO de $\hat{\beta}_0$ et $\hat{\beta}_1$ en considérant, de manière itérative, différentes valeurs de b_0 et b_1 jusqu'à ce qu'on obtienne celles qui minimisent la somme des carrés des erreurs (les estimateurs des moindres carrés), donnée par l'expression (1.6). Cette méthode serait cependant fastidieuse ; heureusement, il existe une formalisation mathématique, détaillée dans le concept-clé 1.2, qui permet la minimisation de l'équation (1.6) et le calcul des estimateurs des MCO.

1.2.2 Estimation par les MCO de la relation liant le score des tests au ratio élèves-enseignants

À partir des 420 observations de la figure 1.3, on peut estimer par les MCO la droite de régression reliant le ratio élèves-enseignants au score des tests. On obtient une valeur estimée de la pente de $-2{,}28$ et une valeur estimée de la constante de 698,9. Ainsi, pour ces 420 observations, la droite de régression linéaire des MCO est donnée par

$$\widehat{ScoreTest} = 698{,}9 - 2{,}28 \times REE, \tag{1.7}$$

où *ScoreTest* est la moyenne des scores des tests et *REE* est le ratio élèves-enseignants. La notation « ˆ » au-dessus de *ScoreTest* dans l'équation (1.7) se réfère à la valeur estimée obtenue à partir de la droite de régression des MCO, illustrée par la figure 1.3.

La valeur de la pente, −2,28, signifie que l'augmentation du ratio élèves-enseignants d'une unité (d'un seul élève par classe) est associée, en moyenne, à une baisse du score des tests de 2,28 points. Par contre, une baisse de ce même ratio de deux élèves par classe génère, en moyenne, une hausse du score des tests de 4,65 points [$= -2 \times (-2{,}28)$]. La valeur négative de la pente signifie que plus le nombre d'élèves par enseignant est élevé (classes de grand effectif), plus la performance aux tests est faible. En d'autres termes, la variation du score des tests est inversement proportionnelle à la variation du ratio élèves-enseignants.

■ Concept-clé 1.2 Estimateur des MCO, valeurs estimées et résidus

Les estimateurs des MCO [a] de la pente β_1 et de la constante β_0 sont

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{s_{XY}}{s_X^2}, \tag{1.8}$$

où s_{XY} et s_X^2 sont respectivement la covariance entre X et Y et la variance de X, et

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}. \tag{1.9}$$

Les valeurs estimées $\hat{Y}_i$ et des résidus $\hat{u}_i$ des MCO sont

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i,\ i = 1, \ldots, n \tag{1.10}$$

et

$$\hat{u}_i = Y_i - \hat{Y}_i, i = 1, \ldots, n. \tag{1.11}$$

Les estimateurs de la constante ($\hat{\beta}_0$), de la pente ($\hat{\beta}_1$), et du terme résiduel ($\hat{u}_i$), ont été calculés à partir de n observations de X_i et de Y_i, $i = 1, \ldots, n$. Ce sont des estimateurs des valeurs inconnues de β_0, β_1 et du terme d'erreur u_i.

a. Les expressions des estimations des MCO sont démontrées en annexe 1.2. Les expressions des variances et covariances empiriques sont définies dans l'annexe B.

Pour une valeur donnée du ratio élèves-enseignants, il est maintenant possible de prédire les scores des tests par zone scolaire. À titre d'exemple, pour une zone avec des classes de 20 élèves par enseignant, la valeur estimée du score des tests est de $698{,}9 - 2{,}28 \times 20 = 653{,}3$. Cependant, l'absence d'autres déterminants des performances scolaires dans ces zones pourrait altérer la qualité de cette estimation.

La question qui se pose maintenant est de savoir si la valeur de la pente est surestimée ou sous-estimée. Pour répondre à cette question, rappelons que le directeur envisage le lancement d'une campagne de recrutement dans le but de

réduire le ratio élèves-enseignants de 2 élèves. Supposons que sa zone scolaire se situe au niveau de la médiane de l'ensemble des zones. Le tableau 1.1 indique que le ratio médian élèves-enseignants est de 19,7 et le score médian des tests est de 654,5. Suite à une réduction de deux élèves par classe, la valeur de son ratio baissera de deux points (de 19,7 à 17,7), ce qui signifie que son ratio élèves-enseignants passera du 50ème percentile au 10ème percentile. Cette variation est importante et nécessite, par conséquent, le recrutement d'un grand nombre d'enseignants. Comment cette campagne de recrutement affectera-t-elle les scores des tests ?

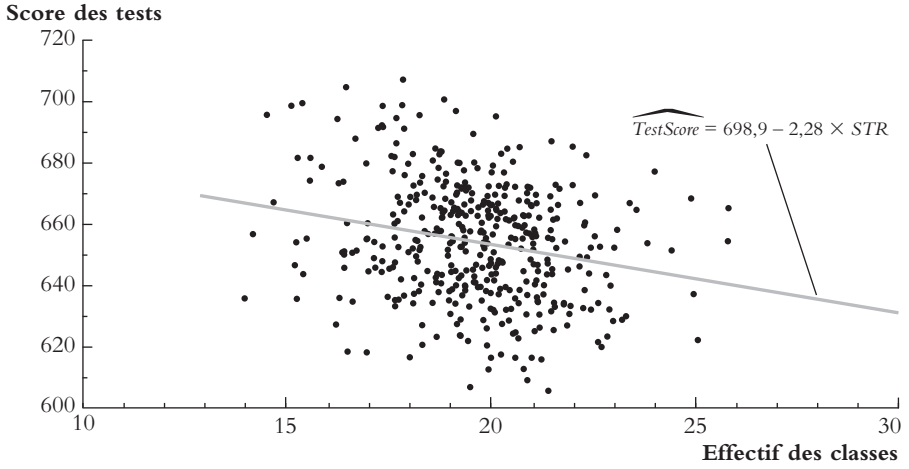

Figure 1.3 Droite de régression estimée

La droite de régression estimée indique l'existence d'une relation négative entre le score des tests et le ratio élèves-enseignants. Une baisse de la taille des classes d'un élève par enseignant conduit à une hausse du score des tests de 2,28 points.

Selon l'équation (1.7), une baisse du ratio élèves-enseignants de 2 élèves par enseignant doit générer une hausse d'approximativement 4,6 points des scores des tests. Si la zone scolaire du directeur est à la médiane, donc avec un score de 654,5, la réduction d'effectif génère une valeur estimée plus élevée de 659,1. Cette hausse est-elle faible ou importante ? Le tableau 1.1 indique qu'une telle variation permettrait à la zone du directeur de se rapprocher du 60ème percentile (ou 6ème décile) des scores. Donc diminuer les classes de 2 élèves placerait cette zone dans une catégorie comportant les 10 % des classes les plus petites, tout en générant le passage de son score aux tests du 5ème décile au 6ème décile. Selon ces estimations du moins, réduire l'effectif des classes d'une manière conséquente (2 élèves par classe), pourrait être bénéfique et profitable, selon la situation budgétaire, mais ce ne sera certainement pas la panacée.

Traitons maintenant les cas extrêmes et évaluons les retombées d'une réduction radicale de 20 élèves par enseignant à 5. Malheureusement, les estimations de

l'équation (1.7) ne seront pas utiles puisque, suivant la figure 1.2, la plus petite valeur du ratio est de 14. Les données utilisées dans cette étude ne contiennent aucune information sur les performances des classes extrêmement petites ; elles ne sont par conséquent pas fiables pour prédire l'effet d'une variation radicale du ratio élèves-enseignants.

1.2.3 Pourquoi utiliser les estimateurs des MCO ?

On utilise les estimateurs des MCO, $\hat{\beta}_0$ et $\hat{\beta}_1$, pour des raisons aussi bien pratiques que théoriques. La méthode des MCO est la plus utilisée dans la pratique ; elle est, en effet, considérée comme le langage commun pour les modèles de régression utilisés dans les domaines de l'économie, de la finance, et plus généralement des sciences sociales. Présenter des résultats à l'aide des MCO (ou de ses variantes, discutées plus tard dans le livre) signifie l'adoption d'un même langage que les autres économistes et que les statisticiens. La formalisation des MCO a été reprise par la quasi-totalité des tableurs et des logiciels de statistiques, ce qui rend facile son utilisation.

À l'instar de l'estimateur de la moyenne théorique d'un échantillon aléatoire, les estimateurs des MCO requièrent des propriétés nécessaires à leur validation théorique. Sous certaines conditions (voir section 1.4), l'estimateur des MCO est sans biais et consistant. Cet estimateur est aussi le plus consistant parmi tous les estimateurs linéaires sans biais. Cependant, cette dernière propriété n'est vérifiée que sous certaines conditions spécifiques qui seront traitées dans la section 2.5.

Encadré 1.1 Le bêta des actifs financiers

Une idée fondamentale de la finance moderne est qu'un investisseur nécessite une incitation financière pour prendre un risque. En d'autres termes, le rendement attendu [1] sur un investissement risqué, R, doit excéder le rendement d'un investissement non risqué, R_f. Ce qui signifie que l'excès de rendement attendu $(R - R_f)$ pour un investissement à risque, tel la détention des obligations d'une compagnie, doit être positif.

Au premier abord, il pourrait sembler que le risque d'une obligation puisse être quantifié par sa variance. Mais une grande partie de ce risque peut être réduite par la détention d'autres obligations dans un portefeuille, c'est-à-dire par sa diversification. Il est donc plus judicieux d'évaluer le risque d'une obligation par sa covariance avec le marché, plutôt que par sa propre variance. Le modèle d'évaluation d'actifs financiers (MEDAF) permet de formaliser cette idée. Selon ce modèle, l'excès de rendement attendu sur un actif financier est proportionnel à l'excès de rendement attendu d'un portefeuille qui contiendrait tous les actifs

1. Le rendement sur un investissement est égal à la variation de son prix plus d'autres gains (dividendes) à partir d'un investissement, exprimé en pourcentage du prix initial. À titre d'exemple, un titre acheté le 1er janvier à 100 €, rapportant durant l'année un dividende de 2,50 € et vendu le 31 décembre à 105 €, devrait avoir un rendement d'un montant $R = [(105 - 100) € + 2{,}50 €]/100 € = 7{,}5\ \%$.

disponibles (« le portefeuille de marché »). Le MEDAF s'écrit

$$R - R_f = \beta(R_m - R_f), \tag{1.12}$$

où R_m est le rendement attendu du portefeuille de marché et β est le coefficient de la régression théorique de $R - R_f$ sur $R_m - R_f$. Dans la pratique, le taux d'intérêt à court terme est souvent considéré comme une bonne approximation (ou *proxy*) du rendement sans risque. Selon le MEDAF, un titre avec $\beta < 1$ est moins risqué que le portefeuille de marché et son excès de rendement attendu sera par conséquent plus faible que celui du marché. Par contre, un titre avec $\beta > 1$ est plus risqué que le portefeuille de marché et génère ainsi un excès de rendement attendu élevé.

Le bêta d'un titre est devenu le cheval de bataille de l'investissement industriel, et les coefficients bêta d'une centaine de titres peuvent être obtenus sur les sites des sociétés d'investissement. Ces bêta sont normalement estimés par la méthode des MCO en régressant l'excès des rendements réels des titres sur l'excès des rendements réels de l'indice du marché.

Le tableau ci-dessous fournit des estimations de bêta pour sept titres français. Les titres des compagnies comme Total et Carrefour, dont les valeurs de bêta sont relativement faibles, sont considérés non risqués. Par contre, les titres de la compagnie d'assurance AXA et de Vivendi sont risqués.

Compagnie	β estimé
Sanofi-Aventis	0,3
Total	0,5
Vinci	0,6
Carrefour	0,7
Renault	1,16
Vivendi	1,50
AXA	1,87

1.3 Évaluation du pouvoir prédictif

Une fois estimée la régression linéaire, reste à évaluer la qualité de l'ajustement du nuage de points par la droite de régression. Cette idée peut être formulée à partir des questions suivantes : la variable indépendante (régresseur) contribue-t-elle fortement à expliquer la variation de la variable dépendante ? Les observations sont-elles regroupées très près de la droite de régression, ou plus dispersées ?

Le R^2 et l'erreur-type de la régression constituent les principales mesures de la qualité d'ajustement des données par la droite de régression. Le R^2 est compris entre 0 et 1 et mesure la part de la variation de Y_i expliquée par X_i. L'erreur-

type de la régression mesure à quel point les valeurs observées Y_i s'écartent de leurs valeurs estimées.

1.3.1 Le R^2

Le **R^2 de la régression** est la part de la variation empirique de Y_i expliquée par (ou estimée par) X_i. Les définitions de la valeur estimée et des résidus (voir concept-clé 1.2) permettent d'exprimer la variable dépendante Y_i comme la somme de la valeur estimée $\hat{Y}_i$ et du terme résiduel $\hat{u}_i$:

$$Y_i = \hat{Y}_i + \hat{u}_i. \tag{1.13}$$

Avec cette notation, le R^2 est le ratio entre la variance empirique de $\hat{Y}_i$ et la variance empirique de Y_i.

Mathématiquement, le R^2 correspond au rapport entre la somme des carrés expliquée et la somme des carrés totale. La **somme des carrés expliquée (SCE)** est la somme des carrés des valeurs estimées de Y_i, $\hat{Y}_i$, centrées par rapport à leur moyenne ; la **somme des carrés totale (STC)**est la somme des carrés des valeurs de Y_i centrées par rapport à leur moyenne :

$$SCE = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \tag{1.14}$$

$$SCT = \sum_{i=1}^{n} (Y_i - \bar{Y})^2. \tag{1.15}$$

L'équation (1.14) est déduite de la propriété selon laquelle la valeur de la moyenne empirique estimée par les MCO est donnée par $\bar{Y}$ (voir la démonstration en annexe 1.3).

Le R^2 est le rapport entre la somme des carrés expliquée et la somme des carrés totale :

$$R^2 = \frac{SCE}{SCT}. \tag{1.16}$$

Le R^2 peut également être exprimé en fonction de la part de variabilité des Y_i qui n'est pas expliquée par X_i. La **somme des carrés des résidus (SCR)** est la somme des carrés des résidus des MCO :

$$SCR = \sum_{i=1}^{n} \hat{u}_i^2. \tag{1.17}$$

Comme démontré en annexe 1.3, *SCT=SCE+SCR*. Donc le R^2 peut aussi être exprimé ainsi :

$$R^2 = 1 - \frac{SCR}{SCT}. \tag{1.18}$$

On observe donc que le R^2 de la régression de Y sur le seul régresseur X est le carré du coefficient de corrélation entre Y et X (la définition du coefficient de corrélation est rappelée en annexe B).

Le R^2 prend ses valeurs entre 0 et 1. Si $\hat{\beta}_1 = 0$, la contribution de X_i dans l'explication de la variation de Y_i est nulle et la valeur estimée de Y_i, fondée sur la régression, est tout simplement sa moyenne empirique. Sous ces conditions, la somme des carrés expliquée (SCE) est nulle ; par conséquent, la somme des carrés des résidus et la somme des carrés totale sont égales et le R^2 sera donc égal à zéro. Au contraire, si la variation de Y_i est totalement expliquée par X_i, nous obtiendrons, pour tout i, la condition $Y_i = \hat{Y}_i$ et le terme résiduel sera par conséquent nul ($\hat{u}_i = 0$). D'où SCT = SCE et $R^2 = 1$. En général, le R^2 n'atteint pas les deux valeurs extrêmes 0 et 1, mais varie entre ces deux valeurs. Un R^2 proche de 1 signifie que le régresseur est un bon prédicteur de Y_i, alors qu'une valeur de R^2 proche de 0 indique que le régresseur est un mauvais prédicteur de Y_i.

1.3.2 L'erreur-type de la régression

L'**erreur-type de la régression (SER)** est un estimateur de l'écart-type de l'erreur de régression u_i. Les unités de mesure de u_i et de y_i étant les mêmes, la SER est une mesure de la dispersion des observations autour de la droite de régression qui exprimée dans la même unité de mesure que la variable dépendante. Par exemple, si l'unité de mesure de la variable dépendante est exprimée en euros, l'ampleur de l'erreur de régression (la SER) sera mesurée en euros.

Puisque les erreurs de régression u_1, $u_2, \ldots,$ u_n sont inobservables, la SER sera calculé à partir de leurs contreparties empiriques, c'est-à-dire les résidus estimés par les MCO, $\hat{u}_1$, $\hat{u}_2, \ldots,$ $\hat{u}_n$. L'expression de la SER est donnée par

$$SER = s_{\hat{u}}, \quad \text{avec} \quad s_{\hat{u}} = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2 = \frac{SCR}{n-2}, \tag{1.19}$$

où dans l'expression de $s_{\hat{u}}$ nous utilisons la propriété selon laquelle la moyenne empirique du terme résiduel des MCO est nulle (voir l'annexe 1.3 pour la démonstration).

La pondération $n-2$ (au lieu de n) de l'équation (1.19) permet de corriger le biais généré par l'estimation des deux coefficients de régression. Ceci est appelé la « correction du degré de liberté » car l'estimation par les MCO de β_0 et β_1 a généré une perte de deux degrés de liberté. Dans l'expression de l'erreur-type, la division par $n-1$ s'explique par la perte d'un degré de liberté suite à l'estimation de la moyenne théorique par la moyenne empirique. Dans le cas où la valeur de n est très grande, retenir les diviseurs, n, $n-1$ ou $n-2$ ne génère aucune modification dans l'expression de l'erreur-type.

1.3.3 Application aux données des scores des tests

L'équation (1.7) correspond à la droite de régression estimée reliant le score des tests standardisés (*ScoreTest*) au ratio élèves-enseignants (REE). Le R^2 de cette régression est de 0,051 ou 5,1 % et la SER est de 18,6.

La valeur de R^2 (0,051) signifie que le régresseur *REE* explique seulement à 5,1 % la variation de la variable dépendante *ScoreTest*. La figure 1.3, qui représente la droite de régression et le nuage de points des données du *ScoreTest* et du *REE*, illustre cette faible contribution prédictive du régresseur : le *REE* explique une partie de la variation du *ScoreTest*, mais une grande partie de cette variation reste à expliquer.

La valeur de la SER, 18,6, correspond à la valeur de l'écart-type des résidus estimés, où l'unité de mesure est le nombre de points (le score) des tests standardisés. Puisque l'écart-type est une mesure de la dispersion, la valeur 18,6 (mesurée en nombre de points des tests) représente un écart important entre le nuage de points et la droite de régression. Cette valeur importante de la SER signifie que la prédiction du scores des tests sur la seule base du régresseur ratio élèves-enseignants va souvent être erronée.

Que peut-on déduire de ce faible R^2 et de ce grand SER ? Ces deux valeurs ne permettent pas de porter un jugement définitif sur la qualité de l'ajustement. Par contre, ce qu'une faible valeur de R^2 nous indique, c'est qu'il existe d'autres facteurs importants qui influencent le score des tests. Ces déterminants pourraient être : les caractéristiques des élèves selon les zones ; le contexte dans lequel les tests se sont déroulés ; les différences de qualité des écoles (autres que celles associées au ratio élèves-enseignants). En résumé, un faible R^2 et un grand SER indiquent une faible contribution du régresseur envisagé, mais ne permettent pas de déterminer les autres facteurs en jeu (contenus dans le terme d'erreur).

1.4 Les hypothèses des moindres carrés

Cette section présente un ensemble de trois hypothèses associées aux modèles de régression linéaire et les conditions statistiques sous lesquelles les MCO fournissent de bons estimateurs des coefficients de régression inconnus β_0 et β_1. Ces hypothèses peuvent paraître abstraites au premier abord, mais elles ont une interprétation aisée. Il est important de bien comprendre le rôle de ces hypothèses, car elles conditionnent la fiabilité et la robustesse des estimations des MCO.

1.4.1 Hypothèse 1 : la distribution conditionnelle de u_i sachant X_i admet une moyenne nulle

La première des trois **hypothèses des moindres carrés** est que l'espérance conditionnelle de u_i sachant X_i est nulle. Cette hypothèse est la formulation mathématique de l'assertion selon laquelle les « autres facteurs » sont contenus dans u_i et non corrélés avec X_i, c'est-à-dire que pour une valeur donnée de X_i, la moyenne de la distribution de ces facteurs est nulle.

Cette hypothèse est illustrée par la figure 1.4 où la droite de régression théorique traduit, en moyenne, la relation entre l'effectif de la classe et les scores des tests. Le terme d'erreur u_i représente l'ensemble des autres facteurs qui génèrent,

pour une zone donnée, l'écart entre les scores des tests et leurs valeurs estimées à partir de la droite de régression théorique. Sur la figure 1.4, on observe que pour une valeur donnée du ratio élèves-enseignants (par exemple 20 élèves par enseignant), ces autres facteurs induisent des performances parfois meilleures ($u_i > 0$), parfois pires ($u_i < 0$) que prédit, mais qu'en moyenne et sur l'ensemble de la population, la prédiction est de bonne qualité. En d'autres termes, pour une valeur donnée $X_i = 20$, la moyenne de la distribution de u_i est égale à zéro. Ce constat pourrait se généraliser pour d'autres valeurs x de X_i, c'est-à-dire que la distribution de u_i, conditionnellement à $X_i = x$, est de moyenne nulle. Mathématiquement, cela peut s'écrire sous la forme $E(u_i|X_i = x) = 0$, ou simplement $E(u_i|X_i) = 0$.

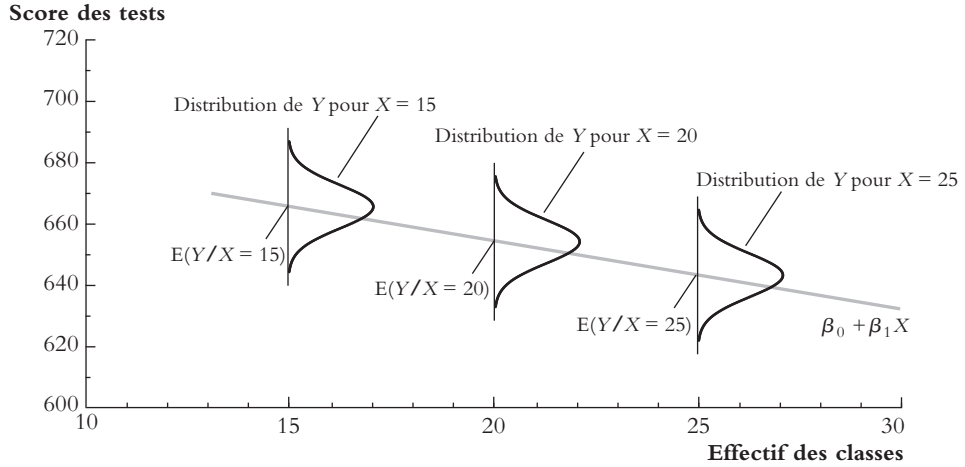

Figure 1.4 Distributions des probabilités conditionnelles et droite de régression théorique

La figure illustre les probabilités conditionnelles du score des tests pour des zones scolaire dont les effectifs des classes sont de 15, 20 et 25 élèves par enseignant. L'espérance de la distribution conditionnelle du score des tests pour un ratio élèves-enseignants donné, $E(Y|X)$, est égale à la droite de régression théorique $\beta_0+\beta_1 X$. Pour une valeur donnée de X, Y est distribuée autour de la droite de régression et le terme d'erreur, $u = Y - (\beta_0 + \beta_1 X)$, admet, pour toute valeur de X, une espérance conditionnelle nulle.

Comme l'indique la figure 1.4, supposer que $E(u_i|X_i) = 0$ revient à supposer que la droite de régression de la population est la moyenne conditionnelle de Y_i sachant X_i (voir la démonstration mathématique de cette propriété dans l'exercice 1.6).

Moyenne conditionnelle de u en expérience aléatoire et idéalement contrôlée

Dans ce type d'expérience, les sujets sont placés aléatoirement dans un groupe de traitement ($X = 1$) ou dans un groupe de contrôle ($X = 0$). L'affectation aléatoire est généralement effectuée par un programme informatique n'utilisant aucune information sur le sujet. Ceci permet d'assurer la condition que la distribution de la variable X est indépendante de l'ensemble des caractéristiques personnelles du sujet. La constitution aléatoire des groupes assure que X et u sont indépendantes ; ce qui implique que la moyenne conditionnelle de u sachant X est nulle.

Dans le cas de données d'observation, X n'est pas aléatoirement assignée dans le cadre d'une expérience contrôlée, et on peut seulement espérer que X se comporte *comme s'*il était assigné aléatoirement, c'est-à-dire que $E(u_i|X_i) = 0$. Vérifier la validité de cette hypothèse pour un cas empirique particulier nécessite une réflexion et une expertise importantes, raison pour laquelle nous reviendrons plusieurs fois sur cet aspect au cours de ce livre (voir en particulier l'annexe B et le chapitre 9).

Corrélation et moyenne conditionnelle

Rappelons que si l'espérance d'une variable aléatoire, conditionnellement à une autre variable, est égale à zéro, les deux variables admettent une covariance nulle et sont par conséquent non corrélées (voir annexe A.3). Donc l'hypothèse selon laquelle $E(u_i|X_i) = 0$ implique que X_i et u_i sont non corrélées ou $cov(u_i, X_i) = 0$. La réciproque n'est pas forcément vérifiée, mais on sait au minimum que si X_i et u_i sont corrélées, alors on a forcément $E(u_i|X_i)$ non nulle. Il est par conséquent plus judicieux de discuter les hypothèses de la moyenne conditionnelle en termes de corrélation possible entre X_i et u_i. Si X_i et u_i sont corrélées, alors l'hypothèse de la moyenne conditionnelle est violée.

1.4.2 Hypothèse 2 : (X_i,Y_i), $i = 1, \ldots, n$ sont indépendamment et identiquement distribuées

La deuxième hypothèse des MCO est que (X_i,Y_i), $i = 1, \ldots, n$ sont indépendamment et identiquement distribuées (i.i.d.). Cette hypothèse concerne la manière avec laquelle l'échantillon est construit. Si les observations sont effectuées à partir d'une seule grande population et par simple échantillonnage aléatoire, (X_i,Y_i), $i = 1, \ldots, n$, seront alors i.i.d. Prenons un exemple. Supposons que X est l'âge d'un employé et Y son salaire, et imaginons qu'on tire quelqu'un au hasard parmi l'ensemble des travailleurs. Cette personne tirée au hasard a un âge et un salaire donnés, donc un X et un Y donnés. Si on tire maintenant aléatoirement n travailleurs de cette population, alors (X_i,Y_i), $i = 1, \ldots, n$ auront nécessairement une distribution identique. À partir d'un tirage aléatoire, (X_i,Y_i), $i = 1, \ldots, n$ sont indépendamment distribués d'une observation à l'autre, c'est-à-dire qu'ils sont i.i.d.

La validité de l'hypothèse d'i.i.d. est plausible dans le cas d'un échantillon comportant un nombre d'observations suffisamment grand. L'hypothèse i.i.d. est plausible pour de nombreuses méthodes de collecte des données. Par exemple, les données d'enquête à partir d'un échantillon tiré au hasard dans la population est un ensemble d'observations i.i.d.

Cependant, tous les plans d'échantillonnage ne produisent pas des observations i.i.d. Par exemple, si les valeurs de X sont fixées par un chercheur dans le cadre d'une expérience, plutôt que tirées aléatoirement à partir d'un échantillon d'une population, alors elles ne seront pas forcément i.i.d. Plus concrètement, intéressons-nous au cas d'un horticulteur souhaitant étudier les effets des techniques bio (X) sur la production de tomates (Y). Il va donc traiter différentes parcelles avec différentes techniques. Si cet horticulteur utilise la même technique bio (niveau de X) sur la $i^{ème}$ parcelle à plusieurs reprises, la valeur de X_i restera inchangée d'un échantillon à l'autre. Donc X_i est non aléatoire (bien que Y_i soit aléatoire), et la méthode d'échantillonnage n'est pas i.i.d. Les résultats présentés dans ce chapitre, concernant les régresseurs i.i.d., sont aussi vérifiés pour les régresseurs non aléatoires. Par exemple, les protocoles expérimentaux modernes incitent l'horticulteur à assigner un niveau de X aux différentes parcelles suivant une répartition aléatoire générée par ordinateur, ce qui contourne un éventuel biais introduit par l'horticulteur (il pourrait utiliser sa technique favorite pour les parcelles de tomates les plus ensoleillées). Quand un tel protocole expérimental moderne est utilisé, le niveau de X est aléatoire et (X_i,Y_i) sont par conséquent i.i.d.

Un autre exemple d'échantillonnage non i.i.d. est le cas des observations se référant à la même unité d'observation au cours du temps. Par exemple, considérons les données sur le niveau des stocks (Y) d'une firme et le taux d'intérêt pratiqué par cette dernière pour ses besoins en liquidité (X), collectées avec une fréquence trimestrielle sur 30 ans. C'est un exemple de série temporelle, dont l'une des propriétés-clés est que les observations temporellement proches les unes des autres ne sont pas indépendantes, mais tendent à être corrélées. Si le taux d'intérêt est faible maintenant, il est probable qu'il le soit le trimestre suivant. Pour ce type de configuration, l'hypothèse i.i.d. n'est plus vérifiée. Les séries temporelles introduisent par conséquent un ensemble de complications qui sera traité plus tard, après la présentation des concepts de base de la régression linéaire.

1.4.3 Hypothèse 3 : les valeurs aberrantes sont rares

La troisième hypothèse des moindres carrés postule que les valeurs extrêmes sont rares, c'est-à-dire que sont peu nombreuses les observations pour lesquelles les valeurs X_i ou Y_i (ou les deux) sont largement en dehors de la concentration normale des données. Les valeurs aberrantes peuvent altérer les résultats des régressions par les MCO. Cette potentielle sensibilité des MCO aux valeurs extrêmes est illustrée par la figure 1.5 utilisant des données hypothétiques.

Dans ce livre, l'hypothèse selon laquelle les valeurs extrêmes sont rares est définie mathématiquement par la condition selon laquelle les moments d'ordre

quatre de X et Y sont non nuls et finis : $0 < E(X_i^4) < \infty$ et $0 < E(Y_i^4) < \infty$. Une autre manière de définir cette hypothèse est de postuler que X et Y admettent une valeur finie du coefficient d'aplatissement (kurtosis, voir annexe A).

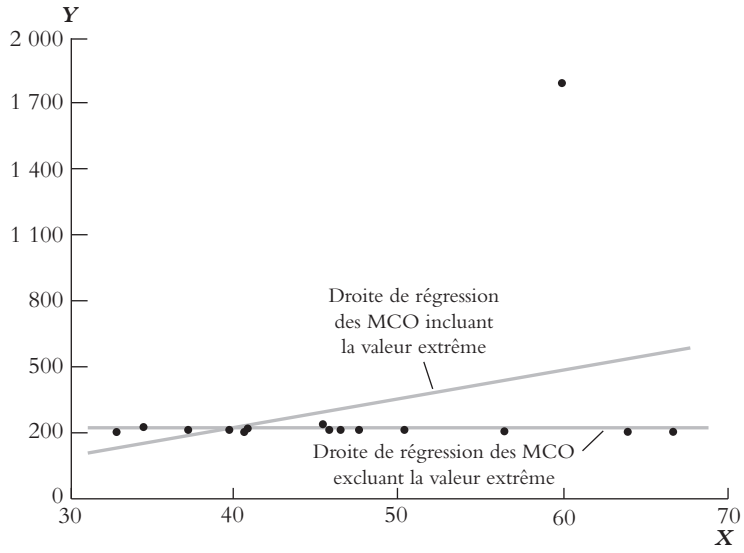

Figure 1.5 Sensibilité des MCO aux valeurs extrêmes
Ces données hypothétiques comportent une valeur extrême. La droite de régression des MCO estimée avec ce point extrême indique une forte relation positive entre X et Y. Cependant, cette relation est quasi inexistante en l'absence du point extrême.

L'hypothèse d'un coefficient d'aplatissement (kurtosis) fini est utilisée pour justifier les approximations asymptotiques (échantillon de grande taille) nécessaires à l'élaboration des distributions des statistiques de tests des MCO. Cette hypothèse est évoquée en annexe B, lors de la discussion des propriétés statistiques de la variance empirique. En particulier, cette hypothèse postule que la variance empirique est un estimateur consistant de la variance théorique. La démonstration de cette propriété nécessite l'utilisation de la loi des grands nombres.

Les erreurs de saisie, telles que les erreurs typographiques ou les erreurs d'unité de mesure, constituent une des sources de l'existence de points extrêmes. Supposons qu'on s'intéresse à la collecte de données concernant la taille des élèves en mètres, et que par erreur, nous ayons enregistré la taille d'un élève en centimètres et non en mètres. Une manière de détecter les points aberrants est de construire les graphiques des données. Si on suppose que la source de ces points aberrants est une erreur de saisie, on peut alors soit la corriger, soit, si c'est impossible, la supprimer.

Mises à part les erreurs de saisie, la validité de l'hypothèse de kurtosis finie est plausible dans la plupart des applications économiques. Dans notre exemple sur des classes de primaire, les effectifs sont plafonnés par la taille matérielle des classes, tandis que les scores des tests dépendent du nombre total possible de réponses justes ou erronées. Ce qui signifie que les données de ces deux variables ont une amplitude finie et par conséquent un kurtosis fini. Plus généralement, les distributions communément utilisées, telle que la distribution normale, admettent des moments d'ordre quatre. Cependant, certaines distributions n'ont pas de moment d'ordre quatre fini, et sont par conséquent exclues de ce cadre d'analyse des régression. Si la troisième hypothèse est vérifiée (il existe un moment fini d'ordre 4), alors il est peu probable que les inférences statistiques basées sur les MCO soient dominées par quelques observations.

1.4.4 Utilisation des hypothèses des MCO

Les trois hypothèses présentées ci-dessus sont résumées dans le concept-clé 1.3. Ces hypothèses jouent deux rôles fondamentaux ; nous y ferons constamment référence dans la suite de ce livre.

■ **Concept-clé 1.3 Hypothèses des moindres carrés**

(hyp. 1) La moyenne conditionnelle du terme d'erreur, u_i sachant X_i, est nulle : $E(u_i|X_i) = 0$.

(hyp. 2) (X_i,Y_i), $i = 1, \ldots, n$ sont indépendamment et identiquement distribués (i.i.d.)

(hyp. 3) Les valeurs aberrantes sont rares : X_i et Y_i admettent des moments d'ordre quatre non nuls et finis.

Leur premier rôle est mathématique : la validité de ces trois hypothèses des moindres carrés conduit à des estimateurs des MCO asymptotiquement distribués suivant une loi normale. Cette propriété asymptotique permet à son tour de développer les tests d'hypothèse et de construire les intervalles de confiance au moyen des estimateurs des MCO.

Leur deuxième rôle est d'établir les conditions permettant la régression par les MCO. Dans la pratique, la première hypothèse des moindres carrés est la plus importante ; elle sera largement discutée dans le chapitre 3 ; un intérêt particulier sera porté aux circonstances conduisant à la violation de cette hypothèse.

Bien que la deuxième hypothèse des moindres carrés soit vérifiée pour la plupart des données transversales, la propriété d'indépendance ne fonctionne pas dans tous les cas, en particulier pour les séries temporelles. Donc les méthodes de régression développées sous l'hypothèse 2 nécessitent une modification dans ces cas.

La troisième hypothèse rappelle que les MCO peuvent être sensibles aux valeurs aberrantes. Si l'ensemble des données contient des valeurs aberrantes, il est

judicieux d'examiner attentivement ces valeurs afin de s'assurer que ces données ont été correctement enregistrées et appartiennent bien à la base de données.

1.5 Distribution des estimateurs des MCO

Comme les estimateurs des MCO, β_0 et β_1 sont estimés à partir d'un échantillon aléatoire, ils sont eux-mêmes aléatoires avec une distribution de probabilité d'échantillonnage décrivant la fréquence d'occurrence de leurs valeurs parmi les possibles échantillons aléatoires. Dans cette section, nous présentons ces distributions empiriques, selon que l'échantillon aléatoire est de petite ou de grande taille. À distance finie (échantillon de petite taille), ces distributions sont compliquées. Par contre, pour de grands échantillons, le théorème central limite permet une approximation asymptotique normale de ces distributions.

1.5.1 Distributions empiriques des estimateurs des MCO

Rappel sur la distribution empirique de $\bar{Y}$

Les sections A.5 et A.6 de l'annexe A portent sur la distribution de l'estimateur $\bar{Y}$ de la moyenne théorique inconnue μ_Y pour de petits et de grands échantillons. Comme les estimateurs des MCO, $\bar{Y}$ est une variable aléatoire qui prend des valeurs différentes d'un échantillon à l'autre, dont les probabilités respectives sont résumées par sa distribution empirique. Bien que, pour un échantillon de petite taille, la distribution empirique de $\bar{Y}$ puisse être compliquée, il est possible d'établir certaines propriétés de cet estimateur qui sont vérifiées pour tout n. En particulier, $\bar{Y}$ est un estimateur sans biais de μ_Y, c'est-à-dire que $E(\bar{Y}) = \mu_Y$. Pour un échantillon de grande taille, le théorème central limite (voir section A.6) établit que cette distribution est approximativement normale.

Distribution asymptotique de $\hat{\beta}_0$ et $\hat{\beta}_1$

La démarche permettant d'évaluer les propriétés statistiques de l'estimateur $\bar{Y}$ peut être étendue aux estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$. En effet, comme pour $\bar{Y}$, $\hat{\beta}_0$ et $\hat{\beta}_1$ sont des variables aléatoires qui prennent différentes valeurs d'un échantillon à l'autre dont les probabilités respectives sont résumées par leurs distributions empiriques.

Bien que pour un échantillon de petite taille, les distributions empiriques de $\hat{\beta}_0$ et $\hat{\beta}_1$ soient compliquées, il est possible d'établir certaines conditions sur ces estimateurs qui sont vérifiées pour tout n. En particulier, sous les hypothèses des moindres carrés ordinaires (voir concept-clé 1.3), nous pouvons écrire :

$$E(\hat{\beta}_0) = \beta_0 \quad \text{et} \quad E(\hat{\beta}_1) = \beta_1. \tag{1.20}$$

Ces deux conditions signifient que les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ sont sans biais. La démonstration que $\hat{\beta}_1$ est sans biais est détaillée en annexe 1.3 et la démonstration que $\hat{\beta}_0$ est sans biais fait l'objet de l'exercice 1.7. Pour un échantillon

de grande taille (asymptotiquement) et suivant le théorème central limite (voir annexe A.4), la distribution d'échantillonnage de $\hat{\beta}_0$ et $\hat{\beta}_1$ peuvent être approchées par une distribution normale bivariée. Ceci implique que les distributions marginales de $\hat{\beta}_0$ et $\hat{\beta}_1$ sont asymptotiquement normales.

Cet argument invoque le théorème central limite. Techniquement, le théorème central limite concerne la distribution d'échantillonnage des moyennes (comme $\bar{Y}$). Si on examine le numérateur de l'expression de $\hat{\beta}_1$ donnée par l'équation (1.8), on peut remarquer que c'est aussi une sorte de moyenne : pas une moyenne simple, comme $\bar{Y}$, mais la moyenne du produit $(Y_i - \bar{Y})(X_i - \bar{X})$. Comme discuté dans l'annexe 1.3, le théorème central limite s'applique à ce type de moyenne. Ainsi, à l'instar de la moyenne simple $\bar{Y}$, cette moyenne est asymptotiquement distribuée suivant une loi normale.

Pour un nombre d'observations suffisamment grand, cette approximation normale des estimateurs des MCO est résumée dans le concept-clé 1.4 (l'annexe 1.3 contient la démonstration de ces formules). Une question pertinente, souvent posée dans la pratique, concerne la valeur appropriée de n pour que cette approximation soit fiable. En annexe A.6, nous suggérons que la valeur $n = 100$ est suffisamment grande pour que la distribution de $\bar{Y}$ puisse être approchée par une distribution normale, même si, parfois un n plus faible suffirait.

■ Concept-clé 1.4 Distributions asymptotiques de $\hat{\beta}_0$ et $\hat{\beta}_1$

Sous les hypothèses des moindres carrés, décrites dans le concept-clé 1.3 et pour un nombre d'observations suffisamment grand, les distributions jointes d'échantillonnage de $\hat{\beta}_0$ et $\hat{\beta}_1$ sont normales. Donc $\hat{\beta}_1$ est asymptotiquement distribuée suivant une loi normale de moyenne β_1 et de variance $\sigma^2_{\hat{\beta}_1}$, $\mathcal{N}(\beta_1, \sigma^2_{\hat{\beta}_1})$, où $\sigma^2_{\hat{\beta}_1}$ est donné par

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n}\frac{var\left((X_i - \mu_X)u_i\right)}{\left(var(X_i)\right)^2}. \tag{1.21}$$

La distribution asymptotique normale de $\hat{\beta}_0$ est $\mathcal{N}(\beta_0, \sigma^2_{\hat{\beta}_0})$, où

$$\sigma^2_{\hat{\beta}_0} = \frac{1}{n}\frac{var\left(H_i u_i\right)}{\left(E\left(H_i^2\right)\right)^2} \quad \text{avec} \quad H_i = 1 - \left(\frac{\mu_X}{E\left(X_i^2\right)}\right) X_i. \tag{1.22}$$

Dans quasiment toutes les applications économétriques modernes, $n > 100$, donc suffisamment grand pour que la distribution d'échantillonnage de $\bar{Y}$ soit approchée par une distribution normale.

Le résultat du concept-clé 1.4 implique que les estimateurs des MCO sont consistants, c'est-à-dire que pour une taille d'échantillon suffisamment grande, $\hat{\beta}_0$ et $\hat{\beta}_1$ sont proches des vraies valeurs théoriques des paramètres inconnus β_0 et β_1, avec une probabilité élevée. Ceci pourrait être expliqué par le fait que les

variances, $\sigma^2_{\hat{\beta}_0}$ et $\sigma^2_{\hat{\beta}_1}$, décroissent vers zéro quand n croît (n apparaît dans le numérateur des expressions des variances). Ainsi, pour n suffisamment grand, les distributions des estimateurs des MCO sont étroitement concentrées autour de leurs moyennes β_0 et β_1.

Les propriétés décrites par le concept-clé 1.4 impliquent que la variance de $\hat{\beta}_1$, $\sigma^2_{\hat{\beta}_1}$, est inversement proportionnelle à la variance de X_i. Pour mieux comprendre cette propriété, analysons la figure 1.6 représentant un nuage de 150 points artificiels de X et Y. Les observations, indiquées par des points gris, correspondent aux 75 points les plus proches de $\bar{X}$. Ainsi, chercher à tracer la droite la plus proche possible des points revient à arbitrer entre les points noirs et les points gris ; personnellement, lesquels choisiriez-vous ? Il semble plus commode de choisir le nuage de points noirs, dont la variance est plus grande, plutôt que le nuage de points gris. De manière similaire, plus la variance de X est importante, plus le paramètre $\hat{\beta}_1$ est précis.

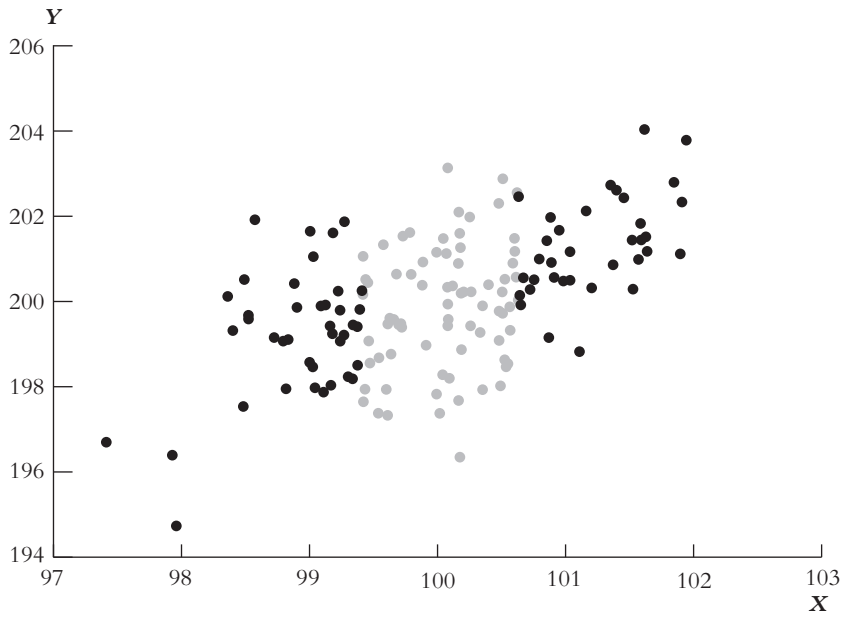

Figure 1.6 Variance de β_1 et variance de X
Les points gris correspondent à un ensemble de X avec une faible variance. Les points noirs correspondent à un ensemble de X avec une large variance. La droite de régression est estimée plus précisément à l'aide des points noirs que des points gris.

Une autre information ressort du concept-clé 1.4 : plus la variance de u_i est faible, plus la variance de $\hat{\beta}_1$ l'est. Cette propriété peut être démontrée mathématiquement. En effet, u_i figure seulement au numérateur de l'équation (1.21) ; ainsi, pour des valeurs données de X, si toutes le valeurs de u_i baissent de

moitié, $\sigma_{\hat{\beta}_1}$ et $\sigma^2_{\hat{\beta}_1}$ baissent respectivement de moitié et d'un quart (voir exercice 1.13). Dit plus simplement, plus les valeurs du terme d'erreur sont faibles (à valeur fixe de X), plus la droite de régression théorique passera près du nuage des points, et sa pente sera donc estimée de manière plus précise. L'approximation normale des distributions asymptotiques de $\hat{\beta}_0$ et $\hat{\beta}_1$ est un outil statistique puissant. En effet, avec cette approximation, il est possible de développer des méthodes d'inférence statistique pour les vraies valeurs théoriques des paramètres de la régression, en utilisant seulement un échantillon.

1.6 Conclusion

Ce chapitre a été consacré à la présentation de la méthode des moindres carrés ordinaires. Le but est d'estimer, à partir d'un échantillon de n observations, la constante et la pente de la droite de régression d'une variable dépendante Y en fonction d'un seul régresseur X, pour une population. Il existe plusieurs manières d'ajuster le nuage de points par une droite de régression ; la méthode des MCO demeure, cependant, la plus utilisée. Lorsque les hypothèses des moindres carrés sont vérifiées, les estimateurs des MCO de la constante et de la pente sont sans biais et consistants, et les variances de leurs distributions empiriques sont inversement proportionnelles à l'effectif de l'échantillon, n. En outre, si n est grand, alors la distribution des estimateurs des MCO est normale.

Ces propriétés fondamentales concernant la distribution des estimateurs des MCO sont vérifiées sous les trois hypothèses des moindres carrés.

La première hypothèse postule que dans le modèle de régression linéaire, la moyenne conditionnelle du terme d'erreur (sachant le régresseur X) est nulle. Cette propriété implique que l'estimateur des MCO est sans biais.

La deuxième hypothèse postule que (X_i,Y_i), $i = 1, \ldots, n$ sont i.i.d. ; c'est notamment le cas des données collectées à partir d'un simple échantillon aléatoire.

La troisième hypothèse concerne la présence de points extrêmes, lesquels sont considérés aberrants. Plus précisément, cette hypothèse signifie que les moments d'ordre quatre des variables X et Y sont finis (kurtosis fini). Cette hypothèse est nécessaire, car sinon, la méthode des MCO pourrait être non fiable en raison de nombreuses valeurs aberrantes. Lorsque ces trois hypothèses sont validées, l'estimateur des MCO est asymptotiquement normalement distribué, comme nous l'indique le concept-clé 1.4.

Ce chapitre établit le cadre théorique des distributions d'échantillonnage des estimateurs des MCO. Il n'est cependant pas suffisant pour développer les tests d'hypothèse ou les intervalles de confiance concernant les valeurs de β_1. Ceci nécessite, en effet, l'estimation de l'écart-type de la distribution d'échantillonnage, c'est-à-dire l'erreur-type de l'estimateur des MCO. Le passage de la distribution de $\hat{\beta}_1$ à l'estimation de son écart-type et à la construction des hypothèses des tests et de l'intervalle de confiance feront l'objet du chapitre suivant.

Résumé

1. La droite de régression théorique, $\beta_0+\beta_1 X$ correspond à la moyenne de Y en fonction de la valeur X. La pente β_1 représente la variation attendue de Y, suite à une variation unitaire de X. La constante β_0 détermine le niveau (hauteur) de la droite de régression. Le concept-clé 1.1 résume la terminologie associée au modèle de régression linéaire.
2. La droite de régression théorique peut être estimée par la méthode des MCO en utilisant les observations (X_i,Y_i), $i = 1, \ldots, n$. Les estimateurs des MCO de la constante et de la pente de la régression sont notés $\hat{\beta}_0$ et $\hat{\beta}_1$.
3. Le R^2 et l'erreur-type de la régression (SER) évaluent la qualité d'ajustement de Y_i par la droite de régression estimée. La valeur de R^2 est comprise entre 0 et 1 ; une grande valeur de R^2 indique que l'ajustement est de bonne qualité. L'erreur-type de la régression est un estimateur de l'écart-type de l'erreur de régression.
4. Il existe trois hypothèses-clés pour le modèle de régression linéaire : (1) la moyenne conditionnelle de l'erreur de régression, u_i, conditionnellement à des valeurs des régresseurs, X_i, est nulle, (2) les observations de l'échantillon sont i.i.d., tirées aléatoirement à partir de la population, et (3) les valeurs aberrantes sont rares. La validité de ces hypothèses conduit à des estimateurs des MCO de β_0 et β_1 qui sont (1) sans biais, (2) consistants et (3) et asymptotiquement normalement distribués.

1.7 Questions et exercices

Révision des concepts

1.1 Expliquez la différence entre β_0 et β_1 ; entre le terme résiduel $\hat{u}_i$ et l'erreur de la régression u_i ; entre les valeurs estimées des MCO, $\hat{Y}_i$ et Y_i.

1.2 Pour chaque hypothèse des MCO, fournissez un exemple pour lequel l'hypothèse est validée et un exemple pour lequel l'hypothèse est violée.

1.3 Dessinez un nuage de points hypothétiques pour une régression correspondant aux valeurs de $R^2 = 0{,}9$ et $0{,}5$.

Exercices

1.1 Supposons qu'un chercheur se propose d'estimer la régression du score des tests sur l'effectif des classes (CS). Les résultats des estimations pour 100 classes sont donnés par

$$\widehat{ScoreTest} = \underset{(20,4)}{520{,}4} - \underset{(2,21)}{5{,}82} \times CS, \quad R^2 = 0{,}08, \quad SER = 11{,}5,$$

où les termes entre parenthèses sont les erreur-types.

a. Déterminez la valeur attendue du score des tests pour une classe de 22 élèves.

b. Supposons que la classe comportait 19 élèves l'année dernière et 23 élèves cette année. Quelle est la variation attendue du score des tests induite par cette variation de l'effectif de la classe ?

c. Pour un échantillon de 100 classes, quel est le score des tests moyen associé à une taille moyenne des classes de 21,4 (utilisez la formule de l'estimateur des MCO) ?

d. Quelle est la valeur de l'écart-type du score des tests pour les 100 classes de l'échantillon (utilisez les expressions de R^2 et de la SER) ?

1.2 Considérons un échantillon aléatoire de 200 personnes de sexe masculin et âgés de 22 ans, dont le poids et l'effectif ont été recensés. L'estimation de la régression de l'effectif en fonction du poids nous donne

$$\widehat{Poids} = \underset{(2,15)}{-99,41} + \underset{(0,31)}{3,94} \times Taille, \quad R^2 = 0,81, \quad SER = 10,2,$$

où *Poids* et *Taille* sont mesurés, respectivement, en livres et en pouces.

a. Quelle sera la valeur estimée du poids d'un homme de 70 pouces ? de 65 pouces ? de 74 pouces ?

b. Un des hommes a une poussée de croissance tardive et grandit de 1,5 pouces au cours de l'année. Quelle sera la variation attendue de son poids induite par son changement de taille ?

1.3 Considérons un échantillon aléatoire composé de salariés diplômés de l'université, âgés entre 25 et 26 ans. L'estimation, à partir de cet échantillon, de la régression du salaire hebdomadaire moyen (*SHEM*, mesuré en €) par rapport à l'âge (mesuré en années) fournit :

$$\widehat{SHEM} = 696,7 + 9,6 \times Âge, \quad R^2 = 0,023, \quad SER = 624,1.$$

a. Explicitez la signification des valeurs des paramètres estimés, 696,7 et 9,6.

b. L'erreur-type de la régression (SER) est de 624,1. Quelle est l'unité de mesure de la SER (€, années, sans unité) ?

c. Le R^2 de la régression est égal à 0,023. Quelle est son unité de mesure (€, années, sans unité) ?

d. À partir des résultats de l'estimation, calculez le salaire d'un cadre âgé de 25 ans ; de 45 ans.

e. D'après ce que vous savez sur la distribution des salaires, pensez-vous que l'hypothèse de normalité de la distribution du terme d'erreur de cette régression est plausible ? Pensez-vous que la distribution est symétrique ou leptokurtique (coefficient d'aplatissement supérieur à 3) ? Quelle est la valeur minimale des salaires ? Est-elle compatible avec la distribution normale ?

f. Pour cet échantillon, l'âge moyen est de 41,6 années. Quelle est la valeur moyenne de *SHEM* pour cet échantillon (utilisez le concept-clé 1.2) ?

1.4 Cet exercice utilise le contenu de l'encadré 1.1 de la section 1.2.

a. Supposons que pour un titre donné, la valeur de β est supérieure à un. Montrez que pour ce titre, la variance de $(R-R_f)$ est supérieure à la variance de $(R_m - R_f)$.

b. Supposons que pour un titre donné, la valeur de β est inférieure à un. Serait-il possible que la variance de $(R - R_f)$ soit supérieure à la variance de $(R_m - R_f)$ (n'oubliez pas de prendre en compte le terme d'erreur de la régression) ?

c. Pour une année donnée, le taux d'intérêt à 3 mois est de 3,5 % et le taux de rendement sur un portefeuille largement diversifié est de 7,3 %. Pour chaque compagnie listée dans l'encadré, utilisez la valeur estimée de β pour évaluer le taux de rendement attendu des titres.

1.5 Un professeur se propose d'évaluer l'impact des différentes conditions d'examen sur la note finale. Les 400 élèves de l'échantillon auront le même sujet d'examen, mais certains auront 30 minutes de moins pour composer (90 minutes pour certains élèves et 120 minutes pour d'autres). Chaque élève est aléatoirement assigné à l'un de ces deux groupes. Supposons que Y_i désigne la note à l'examen final de l'élève i $(0 \leq Y_i \leq 100)$; supposons que X_i désigne la durée durant laquelle l'élève i a composé ($X_i = 90$ ou 120). Considérons, enfin, que le modèle de régression est donnée par $Y_i = \beta_0 + \beta_1 X_i + u_i$.

a. Explicitez la signification du terme d'erreur. Expliquez pourquoi les élèves devraient avoir différentes valeurs de u_i.

b. Expliquez pourquoi, pour ce modèle de régression, $E(u_i|X_i) = 0$.

c. Les hypothèses du concept-clé 1.3 sont-elles vérifiées ? Argumentez.

d. La régression estimée est donnée par $\hat{Y}_i = 49 + 0{,}24X_i$. Calculez le score moyen estimé pour une durée d'examen de 90 minutes. Pour une durée de 120 minutes. Pour une durée de 150 minutes.

1.6 Montrez que la première hypothèse des moindres carrés, $E(u_i|X_i) = 0$ implique que $E(Y_i|X_i) = \beta_0 + \beta_1 X_i$.

1.7 Montrez que $\hat{\beta}_0$ est un estimateur sans biais de β_0 (utilisez le résultat que $\hat{\beta}_1$ est sans biais, démontré dans l'annexe 1.3).

1.8 Supposons qu'à l'exception de la première hypothèse remplacée par $E(u_i|X_i) = 2$, toutes les hypothèses de régression décrites dans le concept-clé 1.3 sont satisfaites. Précisez quels résultats du concept-clé 1.4 sont toujours vérifiés. Argumentez : $\hat{\beta}_1$ est-il asymptotiquement distribué suivant une loi normale de moyenne et de variance données dans le concept-clé 1.4 ? Qu'en est-il de $\hat{\beta}_0$?

1.9 a. Pour une régression linéaire fournissant un $\hat{\beta}_1 = 0$, montrez que $R^2 = 0$.

b. Une régression linéaire avec un $R^2 = 0$ implique-t-elle automatiquement $\hat{\beta}_1 = 0$?

1.10 Supposons que $Y_i = \beta_0 + \beta_1 X_i + u_i$, où (X_i,u_i) sont i.i.d., et X_i est une variable aléatoire suivant une loi de Bernoulli avec $P(X = 1) = 0{,}20$. Pour $X_i = 1$, u_i suit un loi $\mathcal{N}(0{,}4)$ alors que pour $X_i = 0$, u_i est $\mathcal{N}(0{,}1)$.

a. Montrez que les hypothèses décrites par le concept-clé 1.3, sont satisfaites.

b. Déduisez l'expression asymptotique de la variance de $\hat{\beta}_1$ [suggestion : évaluez les termes de l'équation (1.21)].

1.11 Considérons le modèle de régression $Y_i = \beta_0 + \beta_1 X_i + u_i$.

a. Supposons connue la valeur de $\beta_0 = 0$. Déduire l'estimateur des moindres carrés de β_1.

b. Supposons connue la valeur de $\beta_0 = 4$. Déduire l'estimateur des moindres carrés de β_1.

1.12

a. Montrez que le R^2 de la régression de Y sur X est le carré de la corrélation empirique entre X et Y, ce qui revient à montrer que $R^2 = r_{XY}^2$.

b. Montrez que le R^2 de la régression de Y sur X est le même que le R^2 de la régression de X sur Y.

c. Montrez que $\beta_1 = r_{XY}(s_Y/s_X)$, où r_{XY} est la corrélation empirique entre X et Y, avec s_X et s_Y correspondant aux écarts-types empiriques de X et de Y.

1.13 Supposons que $Y_i = \beta_0 + \beta_1 X_i + \kappa u_i$, où κ est une constante non nulle et (X_i,Y_i) satisfont les trois hypothèses des moindres carrés. Montrez que la variance asymptotique de $\hat{\beta}_1$ est donnée par $\sigma^2_{\hat{\beta}_1} = \kappa^2 \frac{1}{n} \frac{var[(X_i - \mu_X)u_i]}{[var(X_i)]^2}$. (Cette équation correspond à la variance, donnée par l'équation (1.21), multipliée par κ^2.)

1.14 Montrez que la droite de régression empirique passe par le point $(\bar{X},\bar{Y})$.

Exercices empiriques

1.1 Pour effectuer cet exercice empirique, reportez-vous à la base de données **CPS08** du site en ligne **http://wps.pearson.fr/principes_econometrie_3**. Ce fichier contient les données, pour l'année 2008, portant sur des salariés âgés de 25 à 34 ans et diplômés de l'université.

a. Estimez la régression du salaire horaire moyen (AHE) sur l'âge (Age). Donnez les valeurs estimées de la constante et de la pente. Selon les résultats de l'estimation, de combien progresse le salaire d'un cadre âgé d'une année de plus ?

b. En utilisant les résultats de l'estimation, déterminez le salaire horaire moyen d'un cadre âgé de 26 ans ; d'un cadre âgé de 30 ans.

c. L'âge est-il un facteur prépondérant de la variation des salaires entre les individus ?

1.2 Sur le même site **http://wps.pearson.fr/principes_econometrie_3**, vous trouverez un fichier appelé **Collegedistance** (« Distance au lycée ») et contenant des données tirées d'un échantillon aléatoire des anciens lycéens qui ont été interviewés en 1980 et réinterviewés en 1986. Le but de cet exercice est d'utiliser ces données pour évaluer la relation entre le

nombre d'années d'études, pour un jeune adulte, et la distance qui sépare son lycée de son université (plus l'université est proche, moins le coût des études est élevé).

a. Estimez la régression des années d'études (ED) sur la distance à la plus proche université ($Dist$), où $Dist$ est mesurée en km (par exemple, $Dist = 20$ signifie que la distance est de 20 km). Donnez les valeurs de la constante et de la pente estimées. D'après les résultats des estimations, déterminez le nombre d'années moyen d'études lorsque que le lycée fréquenté par le jeune adulte est juste à côté de l'université.

b. Quel est le nombre moyen d'années d'études d'un élève fréquentant un lycée à 20 km de l'université ? à 10 km ?

c. La distance séparant le lycée de l'université est-il un facteur prépondérant des différences de niveau d'éducation entre les élèves ?

d. Quelle est la valeur de l'erreur-type de la régression ? Quelle est l'unité de l'erreur-type (km, années, ...) ?

1.3 Sur le même site **http://wps.pearson.fr/principes_econometrie_3**, il y a un fichier appelé **Growth** (« Croissance ») contenant des données, pour 65 pays, sur le taux de croissance entre 1960 et 1995. Le but de cet exercice est d'utiliser ces données pour évaluer la relation entre le taux de croissance et les termes d'échange entre les pays [2].

a. Construisez le nuage de points associé au taux de croissance annuel (*Growth*) et aux termes d'échange (*TradeShare*, c'est-à-dire « Commerce International »). Suivant ce nuage de points, la relation entre ces deux variables est-elle apparente ?

b. Malte est un pays dont les termes d'échange sont plus importants que pour les autres pays. Détectez ce pays sur le nuage de points. Peut-on dire que Malte correspond à un point extrême ou aberrant ?

c. Utilisez l'ensemble des données et estimez la régression de *Growth* sur *TradeShare*. Donnez les valeurs estimées de la constante et de la pente. Grâce aux résultats de l'estimation, déterminez, pour un pays donné, les taux de croissance estimés correspondant à un terme d'échange de 0,5 ; de 1,0.

d. Réestimez le modèle de régression en excluant Malte de la base de données. Répondez de nouveau aux questions c.

2. Cette étude, intitulée « Finance and the Sources of Croissance », a été élaborée par le professeur Ross Levine de l'université de Brown et a été publiée dans *Journal of Financial Economics*, 2000, 58 : 261-300.

Annexes du chapitre 1

Annexe 1.1 La base de données des scores des tests de Californie

La base de données concernant les tests scolaires standardisés en Californie comporte des données sur les performances aux tests, les caractéristiques des écoles et les spécificités démographiques des élèves. Les données utilisées dans ce livre proviennent des 420 zones scolaires californiennes K-6 et K-8 pour l'année 1999. Les scores des tests sont les moyennes des notes en lecture et en mathématiques, pour le test Standard 9, un test administré aux élèves de la classe 5ème. Les caractéristiques des écoles (en moyenne, par zone scolaire) concernent les inscriptions, le nombre d'enseignants titulaires à plein temps, le nombre d'ordinateurs par classe et les dépenses par élève. Le ratio élèves-enseignants utilisé dans ce livre correspond au nombre d'élèves dans la zone scolaire divisé par le nombre d'enseignants. Les variables démographiques pour les élèves sont aussi moyennées par zone scolaire. Ces variables comportent le pourcentage d'élèves faisant partie du programme d'assistance publique, le taux d'élèves qualifiés pour bénéficier des tickets de cantine à prix réduit et le pourcentage d'élèves pour lesquels l'anglais est la deuxième langue. Toutes ces données ont été obtenues à partir du département californien de l'éducation (www.cde.ca.gov).

Annexe 1.2 Démonstration des expressions des estimateurs des MCO

Dans cette annexe, nous allons montrer mathématiquement les formules correspondant aux estimateurs des MCO, données par le concept-clé 1.2. Pour minimiser la somme des carrés des erreurs de prédiction $\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2$ [voir équation (1.2)], nous considérons dans un premier temps les dérivées de premier ordre pour b_0 et b_1 ; elles sont données par

$$\frac{\delta}{\delta b_0}\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2 = -2\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i) \tag{1.23}$$

et

$$\frac{\delta}{\delta b_1}\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2 = -2\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)X_i. \tag{1.24}$$

Les estimateurs des MCO $\hat{\beta}_0$ et $\hat{\beta}_1$ sont les valeurs de b_0 et b_1 minimisant $\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2$; ou, dit autrement, les valeurs de b_0 et b_1 vérifiant les deux conditions du premier ordre, données par les équations (1.23) et (1.24). Sous ces deux conditions, nous pouvons écrire :

$$\bar{Y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{X} = 0 \tag{1.25}$$

et

$$\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \hat{\beta}_0 \bar{X} - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} X_i^2 = 0. \tag{1.26}$$

La résolution de ces deux équations, pour β_0 et β_1, nous donne

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}\bar{Y}}{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{1.27}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}. \tag{1.28}$$

Ces deux équations correspondent aux expressions des estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$, décrites dans le concept-clé 1.2 ; l'expression de $\hat{\beta}_1 = \frac{s_{XY}}{s_X^2}$ est obtenue en divisant le numérateur et le dénominateur de l'équation (1.27) par $n - 1$.

Annexe 1.3 Distributions d'échantillonnage des estimateurs des MCO

Dans cette annexe, nous démontrons que l'estimateur des MCO, $\hat{\beta}_1$, est sans biais et asymptotiquement distribué suivant une loi normale.

Représentation de $\hat{\beta}_1$ en termes de régresseur et d'erreur

Puisque $Y_i = \beta_0 + \beta_1 X_i + u_i$ et $Y_i - \bar{Y} = \beta_1(X_i - \bar{X}) + u_i - \bar{u}$, le numérateur de l'expression de β_1 de l'équation (1.27) s'écrit

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) &= \sum_{i=1}^{n}(X_i - \bar{X})\left(\beta_1(X_i - \bar{X})(u_i - \bar{u})\right) \\ &= \beta_1 \sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u}).\end{aligned} \tag{1.29}$$

On peut écrire : $\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u}) = \sum_{i=1}^{n}(X_i - \bar{X})u_i - \sum_{i=1}^{n}(X_i - \bar{X})\bar{u} = \sum_{i=1}^{n}(X_i - \bar{X})u_i$, qui se déduit de $\sum_{i=1}^{n}(X_i - \bar{X})\bar{u} = \left(\sum_{i=1}^{n}(X_i - n\bar{X}\right)\bar{u} = 0$. En substituant $\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u}) = \sum_{i=1}^{n}(X_i - \bar{X})u_i$ dans l'expression finale de l'équation (1.29), on obtient $\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \beta_1 \sum_{i=1}^{n}(X_i - \bar{X})^2 +$

$\sum_{i=1}^{n}(X_i-\bar{X})u_i$. La substitution de cette expression dans $\hat{\beta}_1$ de l'équation (1.27) fournit :

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}. \tag{1.30}$$

Démonstration de la propriété : $\hat{\beta}_1$ est sans biais

L'expression de $\hat{\beta}_1$ est obtenue en prenant l'espérance de chacun des deux membres de l'équation (1.30) :

$$\begin{aligned} E(\hat{\beta}_1) &= \beta_1 + E\left(\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}\right) \\ &= \beta_1 + E\left(\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})E(u_i|X_1, \ldots, X_n)}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}\right) \qquad (1.31) \\ &= \beta_1, \end{aligned}$$

où la deuxième égalité de l'équation (1.31) résulte de l'utilisation de la loi des espérances itérées (voir annexe A.3). Selon la deuxième hypothèse des moindres carrés, le terme résiduel u_i est indépendamment distribué par rapport à X pour toutes les observations autres que i, d'où : $E(u_i|X_1, \ldots, X_n) = E(u_i|X_i)$. Comme, d'après la première hypothèse des moindres carrés, $E(u_i|X_i) = 0$, l'expression de la moyenne conditionnelle, entre grandes parenthèses dans la deuxième ligne de l'équation (1.31), est nulle et $E(\hat{\beta}_1 - \beta_1|X_1, \ldots, X_n) = 0$. De manière équivalente, $E(\hat{\beta}_1|X_1, \ldots, X_n) = \beta_1$; c'est-à-dire que conditionnellement à $X_1, \ldots, X_n$, $\hat{\beta}_1$ est sans biais. Suivant la loi des projections itératives, $E(\hat{\beta}_1 - \beta_1) = E\left(E(\hat{\beta}_1 - \beta_1|X_1, \ldots, X_n)\right) = 0$, d'où $E(\hat{\beta}_1) = \beta_1$; donc $\hat{\beta}_1$ est sans biais.

Distribution asymptotiquement normale de l'estimateur des MCO

L'approximation asymptotique normale de la distribution de $\hat{\beta}_1$ (voir concept-clé 1.4) est obtenue en considérant le comportement du terme final de l'équation (1.30).

Considérons dans un premier temps le numérateur de ce terme. La consistance de $\bar{X}$ conduit, pour un échantillon de grande taille, à la convergence de $\bar{X}$ vers μ_X. Ainsi, suivant cette convergence, le terme du numérateur de l'équation (1.30) est équivalent à la moyenne empirique $\bar{\nu}$, où $\nu_i = (X_i - \mu_X)u_i$. D'après la première hypothèse des moindres carrés, ν_i est de moyenne nulle. De plus, d'après la deuxième hypothèse des moindres carrés, ν_i est i.i.d. et elle

admet une variance finie et non nulle (troisième hypothèse des moindres carrés) égale à $\sigma_\nu^2 = var[(X_i - \mu_X)u_i]$. Donc $\bar{\nu}$ satisfait les conditions nécessaires à l'application du théorème central limite et $\bar{\nu}/\sigma_{\bar{\nu}}^2$ est par conséquent asymptotiquement distribué suivant une loi $\mathcal{N}(0,1)$, avec $\sigma_{\bar{\nu}}^2 = \sigma_\nu/n$. Il en résulte que la distribution $\bar{\nu}$ peut être approchée par une distribution $\mathcal{N}(0, \sigma_\nu^2/n)$. Considérons, dans un deuxième temps, le dénominateur de l'équation (1.30) : c'est l'expression de la variance empirique de X (divisée par n au lieu de $n-1$, mais ceci est sans conséquence pour un échantillon de grande taille). Suivant la loi des grands nombres, cette variance empirique converge asymptotiquement vers la variance théorique de X.

La combinaison de ces deux résultats asymptotiques conduit à la validité de la condition $\hat{\beta}_1 - \beta_1 \cong \bar{\nu}/var(X_i)$; donc la distribution d'échantillonnage de $\hat{\beta}_1$ suit $\mathcal{N}(\beta_1, \sigma_{\hat{\beta}_1}^2)$, où $\sigma_{\hat{\beta}_1}^2 = var(\bar{\nu})/\left(var(X_i)\right)^2 = var[(X_i - \mu_X)u_i]/n\left(var(X_i)\right)^2$, ce qui correspond à l'expression de l'équation (1.21).

Éléments algébriques supplémentaires pour les MCO

Les résidus et les valeurs estimées des MCO satisfont :

$$\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i = 0 \tag{1.32}$$

$$\frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i = \bar{Y} \tag{1.33}$$

$$\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i X_i = 0 \quad \text{et} \quad s_{\hat{u}X} = 0 \tag{1.34}$$

$$SCT = SCE + SCR. \tag{1.35}$$

Les équations (1.32)-(1.35) indiquent que les moyennes empiriques des résidus et des valeurs estimées sont respectivement égales à zéro et à $\bar{Y}$; la covariance empirique, $s_{\hat{u}X}$, entre le terme résiduel des MCO et les régresseurs, est nulle ; la somme des carrés totale (SCT) est donnée par la somme des carrés expliquée (SCE) et des résidus (SCR).

Pour démontrer le résultat donné par l'équation (1.32), réécrivons à partir de l'expression de $\hat{\beta}_0$ celle du terme résiduel des MCO : $\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i = (Y_i - \bar{Y}) - \hat{\beta}_1(X_i - \bar{X})$. On a donc

$$\sum_{i=1}^{n}\hat{u}_i = \sum_{i=1}^{n}(Y_i - \bar{Y}) - \hat{\beta}_1\sum_{i=1}^{n}(X_i - \bar{X}).$$

Mais les définitions de $\bar{Y}$ et $\bar{X}$ impliquent que $\sum_{i=1}^{n}(Y_i - \bar{Y}) = 0$ et que $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$. On obtient donc $\sum_{i=1}^{n}\hat{u}_i = 0$.

Pour vérifier l'équation (1.33) , il suffit de noter que $Y_i = \hat{Y}_i + \hat{u}_i$, ce qui implique que $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n}\hat{Y}_i + \sum_{i=1}^{n}\hat{u}_i = \sum_{i=1}^{n}\hat{Y}_i$ (d'après 1.32 pour la dernière égalité).

Pour vérifier l'équation (1.34), notons que, sous la condition $\sum_{i=1}^{n} \hat{u}_i = 0$, $\sum_{i=1}^{n} \hat{u}_i X_i = \sum_{i=1}^{n} \hat{u}_i (X_i - \bar{X})$ et, par conséquent,

$$\begin{aligned}\sum_{i=1}^{n} \hat{u}_i X_i &= \sum_{i=1}^{n} \left((Y_i - \bar{Y}) - \hat{\beta}_1 (X_i - \bar{X}) \right) (X_i - \bar{X}) \\ &= \sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X}) - \hat{\beta}_1 \sum_{i=1}^{n} (X_i - \bar{X})^2 = 0, \end{aligned} \tag{1.36}$$

où l'équation (1.36) est obtenue à partir de l'expression de $\hat{\beta}_1$, donnée par l'équation (1.27). Ce résultat, combiné avec les précédents résultats, implique que $s_{\hat{u}X} = 0$.

L'équation (1.35) peut être déduite à partir des précédents résultats et de quelques développements algébriques :

$$\begin{aligned} SCT &= \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \\ &= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + 2 \sum_{i=1}^{n} (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) \\ &= SCR + SCE + 2 \sum_{i=1}^{n} \hat{u}_i \hat{Y}_i = SCR + SCE, \end{aligned} \tag{1.37}$$

où la dernière égalité découle de $\sum_{i=1}^{n} \hat{u}_i \hat{Y}_i = \sum_{i=1}^{n} \hat{u}_i (\hat{\beta}_0 + \hat{\beta}_1 X_i) = \hat{\beta}_0 \sum_{i=1}^{n} \hat{u}_i + \hat{\beta}_1 \sum_{i=1}^{n} \hat{u}_i X_i = 0$.